Definitions and Explanations of Educator Evaluation Terminology

The following definitions and explanations of educator evaluation terminology are provided:

Data Management System
Evaluation Reform
EVALUATION SYSTEM AND COMPONENTS

- New Jersey State Educator Evaluation System
- District Educator Evaluation Rubrics
    - District Teaching Evaluation Rubric
        - Teaching Practice Evaluation Instrument
        - Competencies
        - Evidence-supported Teaching Practice Evaluation Rubric
        - Research-based Teaching Practice Evaluation Rubric
    - District Principal Evaluation Rubric
        - Principal Practice Evaluation Instrument
        - Evidence-supported Principal Practice Evaluation Instrument
        - Research-based Principal Practice Evaluation Instrument

Evidence
Individual Professional Development Plan
InTASC Model Core Teaching Standards
ISLLC Standards
OBSERVATIONS

- Calibration
- Certification/Proof of Mastery
- Data Capture
- External Observer
- Observation
- Observation Conference
- Observer
- Observer Training
- Inter-rater Agreement

SCHOOL STAFF

- Chief School Administrator
- Supervisor
- Teaching Staff Member

SCORING

- Aspects of Scoring Quality
- Reliability
- Rubric
- Score Drift (Observer Effects)
- Types of Scoring and Quality Control
- Validity

Student Learning Objectives (SLO)
Student Growth Percentiles (SGP)

Summative Rating
Data Management System: In its simplest form, an electronic or Internet-based data system and process for storing, organizing, analyzing and reporting evaluation data.

Evaluation Reform: All activities related to developing, piloting, and implementing new evaluation systems for educators in New Jersey. This work started with the Governor's Educator Effectiveness Task Force in 2011, and has continued with the Excellent Educators for New Jersey (EE4NJ) teacher and principal evaluation pilot programs. The eventual goal is for all New Jersey districts to adopt a rigorous and meaningful educator evaluation system that differentiates between levels of performance and provides feedback for professional support and development.

EVALUATION SYSTEM AND COMPONENTS

- **New Jersey State Educator Evaluation System:** The overarching, integrated system in New Jersey of all processes and components of educator evaluation that are used to generate an annual summative evaluation rating for teaching staff members. This system:
  - Encompasses measures of professional practice and measures of student performance and all aspects of implementation, including training and calibration;
  - Uses four levels of annual summative evaluation ratings;
  - Aligns to professional standards;
  - Links to professional development;
  - Involves District Evaluation Advisory Committees of stakeholders, with prescribed membership; and
  - Includes district educator evaluation rubrics.

- **District Educator Evaluation Rubrics:** The set of criteria, measures, and processes to be used in each district to evaluate educators, including professional practice measures and student performance measures. Each district will have an evaluation rubric specifically for teachers (called a "district teaching evaluation rubric"); another specifically for principals, assistant principals, and vice principals (called a "district principal evaluation rubric"); and evaluation rubrics for other categories of teaching staff members (not yet defined). District educator evaluation rubrics include educator practice evaluation instruments.
  - The **District Teaching Evaluation Rubric** includes:
    - Teaching practice measures
      - Measures assessed by a teaching practice evaluation instrument that includes a scoring guide and is evidence-supported
      - Other measures of teaching practice
    - Student performance measures
      - Student Growth Percentiles
      - Other measures of student performance
  - The **District Principal Evaluation Rubric** includes:
    - Principal practice measures
      - Measures assessed by a principal practice evaluation instrument that includes a scoring guide and is evidence-supported

        - Other measures of principal practice
    - Student performance measures
        - Student Growth Percentiles, High School Proficiency Assessment
        - Other measures of student performance

- **Teaching Practice Evaluation Instrument:** The specific teaching practice tool used to assess the observable competencies of teaching practice. The instrument consists of the rubrics and accompanying definitions and descriptions of the ratings used in assessing teaching practice. It may also include more detailed representations of teaching practice such as indicators or examples. The selected teaching practice evaluation instrument must have an evidence base documenting that it meets various specifications, which are outlined in the NJDOE evaluation FAQs.
    - **Competencies:** The specific indicators of teaching practice that are assessed by a given teaching practice evaluation framework. These may vary between frameworks, but generally they are similar. Some examples include classroom management, questioning, and/or professional responsibility.
    - **Evidence-supported Teaching Practice Evaluation Instrument:** A teaching practice evaluation instrument that provides: (1) scales or dimensions that capture multiple and varied aspects of teaching performance which must be attested by knowledgeable practitioners or experts in the content prior to use in observation of a teacher's practice; (2) differentiation of a range of teaching performance as described by the score scales which must be shown in practice and/or research studies; and (3) objective validation on the aspects of both concurrent and construct validity.
        - Concurrent validity as applied to the instrument means that higher observed instructional quality as measured by the instrument is related to higher student learning achievement or gains. This relationship must be shown through provided data sets or study results.
        - Construct validity as applied to the instrument means that the measure actually assesses the dimension of teaching effectiveness it claims to measure. The establishment of such claim must be attested by knowledgeable practitioners or experts in the content.
    - **Research-based Teaching Practice Evaluation Instrument:** A teaching practice evaluation instrument providing scores or categorizations which have been found to be valid for specified purposes through a research process whereby: (1) studies have been completed using the current form of the instrument that have demonstrated the application of rigorous, systematic, and objective procedures to obtain reliable and valid results; and (2) these results have been published in a format where they have been subject to professional peer review (and preferably blind review).

- **Principal Practice Evaluation Instrument**: A tool used to assess principal practice. The instrument consists of the rubrics and accompanying definitions and descriptions of the scales used in assessing principal practice. It may also include more detailed representations of principal practice such as indicators or examples. The selected

principal practice evaluation instrument must have an evidence base documenting that it meets various specifications, which are outlined in the NJDOE evaluation FAQs.

- **Evidence-Supported Principal Practice Evaluation Instrument**: An evaluation instrument that provides: (1) scales or dimensions that capture multiple and varied aspects of principal performance which must be attested by knowledgeable practitioners or experts in the content prior to use in evaluating a principal's practice; (2) differentiation of a range of principal performances as described by the score scales which must be shown in practice and/or research studies; (3) objective validation on the aspects of both concurrent and construct validity. Concurrent validity as applied to the instrument means that higher observed instructional quality as measured by the instrument is related to higher student learning achievement or gains. This relationship must be shown through provided data sets or study results. Construct validity as applied to the instrument must be attested by knowledgeable practitioners or experts in the content.
- **Research-Based Principal Practice Evaluation Instrument**: An evaluation instrument providing scores or categorizations which have been found to be valid for specified purposes through a research process whereby: (1) studies have been completed using the current form of the instrument that have demonstrated the application of rigorous, systematic, and objective procedures to obtain reliable and valid results; and (2) these results have been published in a format where they have been subject to professional peer review (and preferably blind review).

Evidence: Documents or artifacts that demonstrate or confirm the work of the person being evaluated and support the rating on a given element or component of an evaluation instrument's rubric.

Individual Professional Development Plan: A written statement of actions developed by the supervisor and the teaching staff member to continue the teaching staff member's professional growth and/or correct deficiencies. The individual professional development plan includes timelines for implementation, and responsibilities of the individual teaching staff member and the school district for implementing the plan.

InTASC Model Core Teaching Standards: The 2011 InTASC Model Core Teaching Standards, finalized in May 2011, outline what teachers should know and be able to do to ensure every K-12 student reaches the goal of being ready to enter college or the workforce. These standards were developed in response to the need for a new vision of teaching to meet the needs of next generation learners. These standards outline the common principles and foundations of teaching practice that cut across all subject areas and grade levels and that are necessary to improve student achievement. They are a revision of the 1992 model standards which New Jersey adapted in 2003 as the New Jersey Professional Teaching Standards. At the current time, the 2011 InTASC Model Core Teaching Standards are in the process of being adopted for the purposes of approving and alignment to teacher evaluation. The 2011 standards can be accessed at: http://www.ccsso.org/resources/programs/interstate_teacher_assessment_consortium_(intasc).html.

ISLLC Standards: A set of high-level policy standards to guide education leaders throughout their careers and to inform improvements in education leadership, preparation, licensure, evaluation and professional development.
http://www.ccsso.org/Documents/2008/Educational_Leadership_Policy_Standards_2008.pdf

OBSERVATIONS

- **Calibration**: A process to monitor the scoring of an observer who has been trained and who has demonstrated proof of mastery on a teaching practice evaluation instrument, to ensure that such observer continues to score accurately and consistently according to the standards and definitions of the instrument.

- **Certification/Proof of Mastery:** A set of requirements or assessments used upon completing training to determine whether a trainee observer has achieved mastery of the content of the training as well as accuracy and consistency in using the rubric as applied to practice.

- **Data Capture:** A process by which the data supporting claims associated with the system, such as those related to observer mastery of a rubric, success in calibration, or observation scores and evidence, are captured and stored in a format that can be accessed and used.

- **External Observer:** An individual appropriately trained as an observer and not currently working in the school of the teacher he/she is observing; this observer must be either certified or have demonstrated proof of mastery in the evaluation instrument adopted by the district, and be held to all scoring quality monitoring standards.

- **Observation:** A visit to an assigned work station by an observer for the purpose of formally collecting data on the performance of a teaching staff member's assigned duties and responsibilities and of a duration appropriate to same.

- **Observation Conference:** A discussion between a supervisor and teaching staff member to review a written report of the performance data collected in a formal observation and its implications for the teaching staff member's annual evaluation.

- **Observer:** An individual trained on the Evaluation Instrument as an observer and either certified or demonstrated to have proof of mastery in the teaching Evaluation Instrument adopted by the district, and held to all scoring quality monitoring standards.

- **Observer Training:** The process by which candidate observers learn about the instrument, as well as how to apply accurately and consistently the scales and score levels of the rubric to content that is as similar as possible to that seen in practice.

- **Inter-rater Agreement:** The result when two observers using the same measure to evaluate the same teacher produce the same results in ratings and feedback (sometimes referred to as “inter-rater reliability”). Inter-rater agreement is one aspect considered in

the determination of whether scores from a measure of teaching effectiveness can be considered "reliable." There are some important caveats and conditions when measuring levels of agreement:

- Observers can agree by chance, especially if using rating scales with few score points. There are measures of agreement corrected for chance, such as *kappa*, that help provide a more accurate assessment of what the observers are contributing over and above chance agreement, and *these should be used in preference over raw agreement*.
- Observers can be wrong and agree with each other. Agreement alone does not assure accuracy of scoring—just consistency. Therefore, calibration is necessary to ensure accuracy of scoring.

SCHOOL STAFF

- **Chief School Administrator:** The superintendent of schools, or if there is no superintendent, the administrative principal.

- **Supervisor:** Any appropriately certified individual assigned with the responsibility for the direction and guidance of the work of teaching staff members.

- **Teaching Staff Member:** A member of the professional staff of a school district holding office, position, or employment of such character that the qualifications require him or her to hold a valid and effective standard, provisional, or emergency certificate, issued by the State Board of Examiners.

SCORING

- **Aspects of Scoring Quality:** There are different aspects of scoring quality that are worth defining:
  - *Accuracy* is consistency with master coders—whether the observer assigns the "correct" score to the performance. "Correct" scores must be obtained through a judgment process, most preferably with experts who complete a master-coding process and reach consensus on the final score, evidence, connection with the rubric and score level, and rationale. This aspect is particularly important for observers who may see a limited range of practice (in *any* part of the scale) in their observations. This can lead to "relative scoring" wherein the observed practice scores are spread artificially by the observer to encompass the full score range of the instrument. Observers in such circumstances should be exposed frequently to examples at all levels of practice to reset their scoring to the Evaluation Instrument standards.
  - *Inter-rater agreement* is consistency with other observers—whether two observers completing independent ratings of the same performance agree on the score(s) that they assigned (i.e., two observers using the same measure to evaluate the same teacher produce the same results). This agreement can be exact (no difference in scores), adjacent (usually defined as within one score category of each other's scores), or discrepant (usually defined as more than one score category apart). In

high-stakes situations, it may be necessary to resolve differences in observer scores that are discrepant or even adjacent.
  - *Trend agreement* is consistency over time—whether observers assign the same score to the same performance when scored on occasions separated in time.
  - *Unbiased scoring* is consistency across candidates —whether observers ignore aspects of the performance, teacher, students, teaching style, specific content, setting, or any other facets that are irrelevant to the instrument. Observers improperly influenced in their scoring by such factors should be retrained and recalibrated or removed from scoring.

- **Reliability:** The degree to which an instrument measures something consistently. This measurement property of an instrument must be evaluated across different observers and contexts.
- **Rubric:** A scoring guide composed of criteria used to evaluate performance, a product, or a project. A rubric allows for standardized evaluation according to specified criteria, making scoring and ranking at several levels simpler and more transparent in a reliable, fair, and valid manner.

- **Score Drift (Observer Effects):** Score drift occurs when the scores assigned by an observer to the group of teachers move away from the standard set on the observation rubric. Drift can be positive (scores are more lenient than intended by the instrument) or negative (scores are more stringent than intended by the rubric). Other types of score drift include *scale compression*, when an observer inappropriately uses only part of the scale to assign scores to observations that encompass the entire range of performance, *and scale expansion*, when an observer inappropriately uses the full range of scores on the scale to assign scores to observations occurring in a narrower range of performance. Observers can become more variable (expand their scale) or less variable (compress their scale) over time, even if the range of observed performance remains constant. Observers should be calibrated on a regular basis to ensure that score drift is not occurring. Similarly, quality control measures such as double scoring should also be done on a regular basis to determine if observers' scoring need to be calibrated.

- **Types of Scoring and Quality Control:**
  - Certification and Proof of Mastery are scoring skills assessments completed at the end of training to verify that an observer has learned to apply the rubric accurately. Certification and proof of mastery typically are a relatively extensive assessment of skills and should encompass scoring teaching performance (typically using videos) across the entire score range on all aspects of the rubric so that observers are able to identify what teaching looks like across the scoring continuum.
  - Double-scoring occurs when two (or more) observers assign scores to a performance independently of each other. This can be done by having two observers in the same classroom session or through the use of video capture.

- **Validity:** The degree to which an interpretation of an evaluation score is supported by evidence. For a measure of teaching effectiveness to be valid, evidence must support the

argument that the measure actually assesses the dimension of teaching effectiveness it claims to measure and not something else. Instruments cannot be valid in and of themselves; an instrument or assessment must be validated for particular purposes.

Student Learning Objectives (SLO): A standards-based statement in specific and measurable terms that describes what learners will know or be able to do as a result of mastering the skills and knowledge in the curriculum. As an example, teachers may assess students at the beginning of the year and set objectives, and then assess again at the end of the year (pre- and post-testing). Often the principal or a designee works with teachers to approve the SLO and determine success.

Student Growth Percentiles (SGP): For K-12 education in New Jersey, the phrase "growth model" describes a method of measuring individual student progress on statewide assessments (the NJASK) by tracking student scores from one year to the next. Each student with at least two consecutive years of NJASK scores will receive a *student growth percentile*, which measures how much the student changed relative to other students statewide with similar scores in previous years. SGPs range from 1 to 99, where higher numbers represent higher growth and lower numbers represent lower growth. All students, no matter the scores they earned on past NJASK tests, have an equal chance to demonstrate growth at any of the 99 percentiles on the next year's test. Growth percentiles are calculated in ELA and mathematics for students in grades 4 through 8. Additional SGP information can be found here, and a video tutorial is located here.

Summative Rating: The final annual rating for every teacher, resulting in one of the four following category assignments: highly effective, effective, partially effective, or ineffective. All relevant evaluation data will be combined in a structured way to determine the summative rating.